UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Tarea 4: Agrupamiento

Aprendizaje Automático

Nombre: Sergio Andrés Elizondo Rodríguez

Grupo: 003

Maestro: José Anastacio Hernández Saldaña

Sábado 20 de Julio del 2024

Introducción

NSL-KDD es un conjunto de datos ampliamente utilizado en la investigación de sistemas de detección de intrusiones (*IDS*, por sus siglas en inglés). Está compuesto por registros de tráfico de red, con características que describen cada conexión, y su etiqueta correspondiente, la cual indica si la conexión es normal o se trata de un ataque.

El análisis realizado en el presente reporte busca comparar los modelos de clasificación entrenados anteriormente con modelos de agrupamiento, haciendo uso del algoritmo *k*-medias y optimizando el parámetro *k* de acuerdo con el criterio del codo.

Procesamiento previo

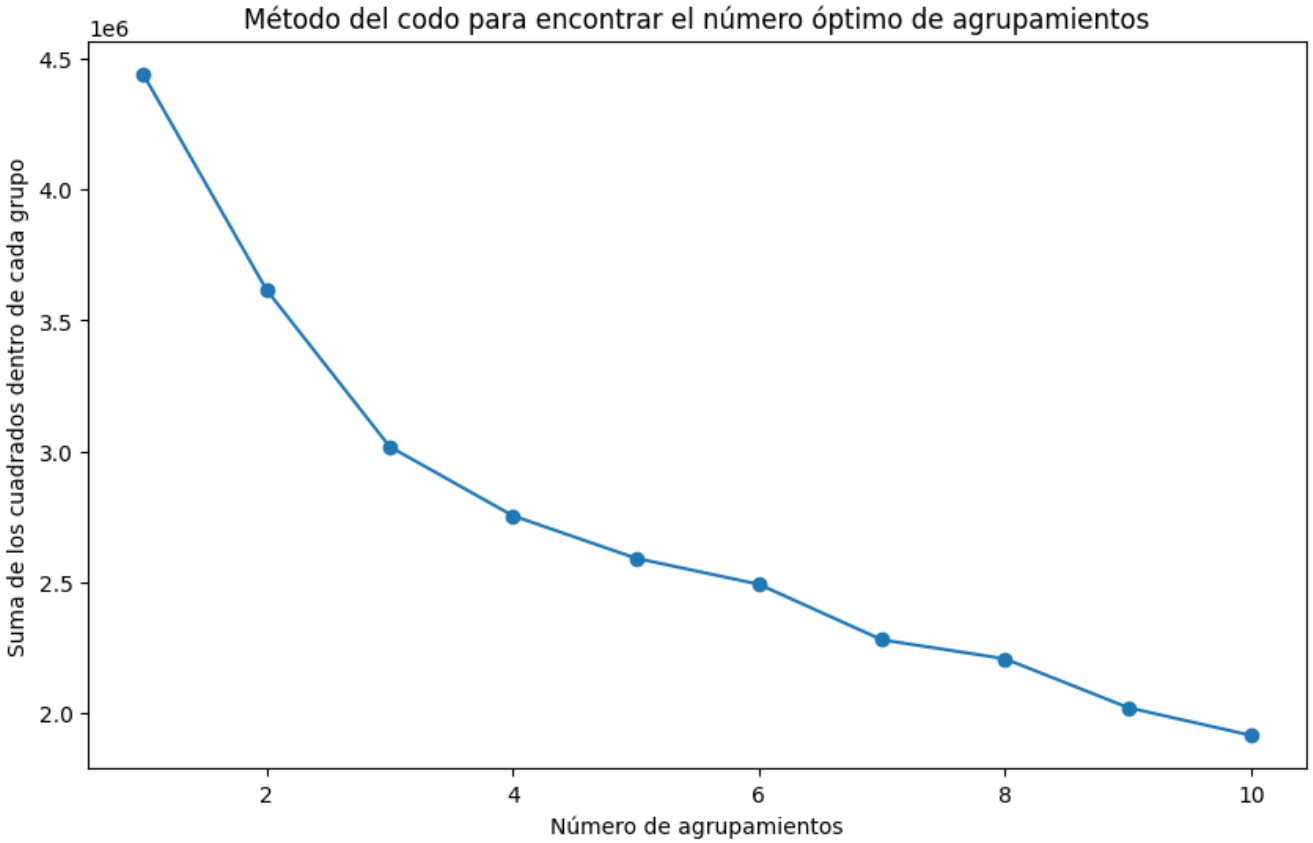
Antes de aplicar los modelos, se realizó un procesamiento previo para garantizar que las variables sometidas a análisis fueran las adecuadas; en particular, se aplicaron dos tratamientos diferentes:

- **Variables categóricas:** Representan categorías o grupos discretos (como el tipo de conexión). Se convirtieron en representaciones binarias utilizando *OneHotEncoder* para poder usarlas en los modelos de regresión.
- **Variables numéricas:** Representan datos cuantitativos (como las métricas de tráfico de red). Se normalizaron para que fueran compatibles con todos los modelos.

Modelos

Se implementaron dos modelos de agrupamiento utilizando el algoritmo *k*-medias, comparando sus resultados con un modelo de clasificación desarrollado previamente.

Primero, se seleccionó $k = 2$, correspondiente al número de clases existentes (normal y anomalía). Posteriormente, se analizó mediante el método del codo el número óptimo de agrupamientos:



Resultados

El número óptimo son 3 agrupamientos. Con el fin de que los resultados sean comparables tanto al primer modelo de *k*-medias como al modelo de clasificación, se asignó la clase correspondiente a cada agrupamiento de acuerdo con las etiquetas existentes (anomalía y comportamiento normal). A continuación, se muestran los resultados de las métricas:

Modelo	***k*-medias (2 agrupamientos)**	***k*-medias (método del codo, 3 agrupamientos)**	**Bosque aleatorio**
Exactitud (*accuracy*)	0.5692	0.5692	0.8038
Precisión (*precision*)	0.5692	0.5692	0.9687
Sensibilidad (*recall*)	1	1	0.6772
Métrica F1	0.7255	0.7255	0.7971
Área bajo la curva *ROC* (*AUC-ROC*)	0.5	0.5	0.8241

Observamos que todas las métricas son mejores para el modelo de clasificación (bosque aleatorio) con excepción de la sensibilidad. Esto indica que los modelos de *k*-medias tienen un mayor sesgo, tendiendo a predecir positivos (lo cual impacta positivamente en la sensibilidad, al no existir falsos negativos, aunque afectando a las demás métricas).

Bibliografía

- **Base de datos:** NSL-KDD. GitHub. Recuperado el 21 de julio de 2024, de https://github.com/topics/nsl-kdd